Frequently asked questions about "A Genetic Atlas of Human Admixture History"

G. Hellenthal, G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, S. Myers, Science (2014)

SUMMARY

What is your work about and how is it new?

Many researchers have studied human DNA and its history, and this has revealed how humans have spread around the globe. This spreading out produces distinct populations, and small genetic differences between separated populations arise (although most variation is still shared among groups). When groups come back together – for example due to migrations or invasions - and have children, this is called genetic admixture, and leaves a characteristic signature in DNA.

Our work uses DNA from many people around the world to identify these mixture events, and find out first, who the groups were that mixed – often to the level of individual countries – and second, when the mixture occurred. Other researchers have developed important tools to look at one or other these questions but our method is the first to do both simultaneously – allowing us to more fully describe events. By using "chromosome painting", it also offers better power and precision than previously available. We are able to consider complex histories (e.g. several waves of mixing) and have analysed 95 groups across the globe, producing an "atlas" of mixing dates, places and mixing populations.

What does your study imply about human ancestry?

(1) Most human populations are a product of mixture of genetically distinct groups that intermixed within the last 4,000 years.

(2) Mixture events are often localized in time and space: neighboring populations can sometimes have distinct ancestry and history, especially for recent events.

(3) Many mixture events involve source populations from very distant locations separated by thousands of miles.

GENERAL QUESTIONS

What is genetic admixture?

Genetic admixture occurs when individuals from two or more genetically distinguishable groups have children together. This might happen when individuals from one part of the world settle into a new geographic region already inhabited by other people, e.g. due to invasions or large-scale migrations.

What was the aim of your study?

Our study uses novel statistical methodology seeks to identify and characterize genetic admixture events in the history of the groups in our sample. Results are presented on an interactive webpage at www.admixturemap.paintmychromosomes.com.

Which data did you use?

The dataset we analyzed here consists of 1,490 individuals sampled from 95 world-wide groups with 2-46 individuals per group. For each individual, we had genotypes at 474,491 genetic markers (Single-Nucleotide-Polymorphisms; SNPs) across all 22 non-sex chromosomes. These data included both new samples and samples collated from multiple publicly available resources.

How have you defined a "group"?

There is a long-standing debate spanning several scientific fields about what constitutes a "group" or "population". For our analysis here, we define a "group" to be a set of individuals whom all have a similar genetic make-up. This is initially based on labels given by researchers during the sampling process (which are dependent on self-identification) but in some cases we refined groups to better reflect genetic similarity using the fineSTRUCTURE algorithm (see below).

What signatures does admixture leave in genetic data?

When individuals from different groups have children (i.e. admix), their offspring's DNA becomes a mixture of the DNA from each admixing group. Pieces of this DNA are passed along through subsequent generations, carrying on all the way to the present day. Therefore, the genomes of modern-day individuals (who descend from this admixed population) contain segments of DNA inherited from each of the original source groups. Figure 1 summarizes this process.

Furthermore, due to a genetic re-shuffling process called "recombination" that occurs each generation, the lengths of uninterrupted DNA segments inherited from each source group reflect how long ago the admixture event occurred. In general if modern-day genomes carry long uninterrupted segments from each source group, the admixture event occurred more recently. In contrast if they carry short segments, the admixture occurred longer ago. We describe how we try to capture this information below.

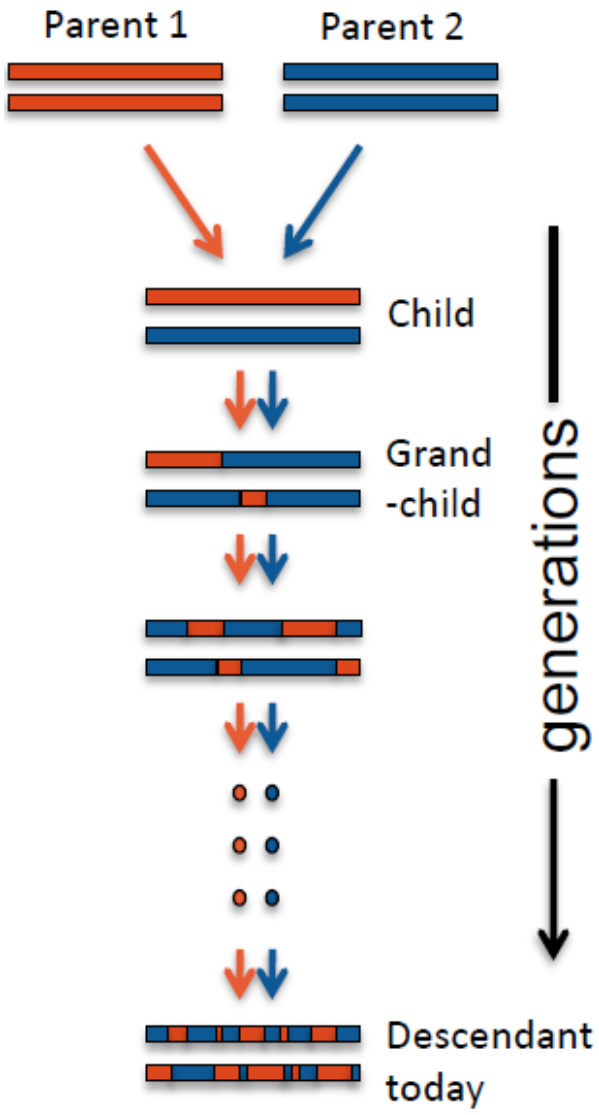

Figure 1: Schematic of the admixture process

How is history, as historians tell it, related to genetic admixture?

Good question! Our approach aims to identify the movements of peoples that resulted in interbreeding or DNA exchange at different periods in human history, and quantify the proportion of DNA transmitted to the present day. Because it does not use other forms of information in doing so, the results can be compared and combined with historical information. Although we have tried to interpret our findings in the light of well known historical events where appropriate, our work concentrates on the genetic analysis and we are not historians! We note that it is extremely difficult to predict the genetic legacy of even well known events – and whether we descend from the architects of these events – without examining DNA directly. This is because the extent to which different events – e.g. empires, or the transmission of materials and cultures – result in genetic mixture occurring is unknown, and deserves further systematic study. The idea of combining history and genetic – to provide exciting insights into both – is one motivation of our work.

In our study, many of the admixture signals we observe do appear to match well, in terms of both times and groups involved, with historical events such as the Arab Slave Trade, Mongol Era expansion and

Slavic/Turkic expansions and help determine the extent and proportion of DNA contributed by each.

Wait a second…..if genetic mixing happened in the past, might the admixing groups have disappeared, or changed beyond recognition, by now? Or what if you didn't sample them? Does your method go wrong in these cases?

This is an important point – it is of course unlikely, for many of the events, that we are truly able to sample the genuine mixing "source" groups. For a given event, it may be that one of the source groups is *close* to a sampled population. In this case (in simulations) our method typically identifies this group as playing a dominating role in the event. For example in the Mozabite (Figure 2) one source (orange) is dominated by the West African "Yoruba" population.

To capture the idea of a source more different to our sampled populations, we allow our inferred sources to consist of a mixture of the sampled populations, not just one. This is shown on the results of the interactive map. In the Mozabite (Figure 2) the second source population is represented as a mixture of two North African groups, suggesting it was likely from North Africa, where the Mozabite live now, but not identical to any of the groups sampled. So we might conclude a North African population and migrants from West Africa mixed, in the ancestors of today's Mozabite people.

A widely spread set of source populations is sometimes seen – especially for ancient events – and suggests that the source population is very distinct from those sampled, or itself had mixed ancestry.

Because mixtures are sometimes hard to interpret, we also provide the single sampled population representing our "best-guess" for the group most similar to the real source, in terms of having the most similar genetic make-up. Especially when the mixture contains many widely spread groups, this best guess is not necessarily the real source, or even from the same place as the real source! Both the inferred mixtures and the best guesses we made in our analyses are provided at www.admixturemap.paintmychromosomes.com.

It is probably also worth mentioning that we believe our *dating* approach is robust to the makeup of the sources – so even in cases where the sources are difficult to interpret, the timing of admixture ought to be correct.

Figure 2: Screenshot from the webpage showing our results regarding genetic mixing in the ancestors of today's Mozabite people. One source group (blue, contributing 92% of the DNA) is fit as a mixture of two sampled groups, from Morocco and Tunisia – implying it is not an exact match for a sampled group. Admixture is dated to 1334CE (1250CE - 1418CE, timeline).

Could things be complicated in other ways? Do people actually come together at some specific time, or could things happen more gradually?

Definitely, things can be complicated. We make an attempt to identify complex cases, by testing whether mixing events happen at one time, or more than one time. We also test whether the number of mixing groups is two, or whether it is larger – say three or more.

Many cases – e.g. all populations we suggest have mixing related to the Mongol empire – do seem to closely fit the idea of a single "pulse" of admixture between two groups. Other populations seem to be affected by mixing with different sources at different times.

However, especially in very complicated cases, we may miss events that have occurred and we are not yet able to analyse, for example, admixture involving more than three groups. We also have limited

ability to completely date admixture in cases where the same, or very similar, groups mix at multiple times in the history of a single descendant group. Two pulses, or a continual "trickle" of migrants, for example, leave very similar traces – even in theory – and we only attempt to infer a date "range" for such cases, leaving them partially described as "admixture at more than one time".

Is there a limit to how far back in time it's possible to look?

With extensive simulations, we found our software GLOBETROTTER to reliably identify, describe and date admixture events occurring <160 generations ago, corresponding to ~4500 years ago. GLOBETROTTER can likely date events somewhat older than this, but this is a rough idea of how old it can go in its current implementation, and technical reasons will make it challenging to go back tens of thousands of years. One very exciting possible way to look back further is to apply the approach to DNA from ancient human remains. Some such individuals are starting to be sequenced by several groups already, and more data will become available in future.

QUESTIONS ABOUT THE RESULTS

Are any signals shared across many human populations?

One event which left a geographically broad and easy to identify signal is the Mongol expansion (7 populations in our sample implicated). A second is the Arab slave trade (18 populations in our sample implicated), with DNA contributions from sub-Saharan Africa inferred to occur at a range of dates over more than a thousand years seen in people from around the Mediterranean (involving migrants from W. Africa), and the Arabian Sea and the Persian gulf (involving migrants from East and southern Africa). The Mongol expansion seems, according to both history and genetics to have been a particularly abrupt transfer of people and DNA across Asia. Events in Eastern Europe also span multiple groups (see below). Many other events appear far more localized, though still very interesting!

Which populations have the most recent mixing in the dataset?

Several groups show very recent admixture, including the American Maya, whose signal of mixing between multiple groups agrees with groups and timing for European (e.g. Spanish) settlers in the region, and West African migration to the Americas through European slavery. European and other DNA in the San Khomani from southern Africa, and Russian-like DNA in the Yakut from Siberia, also originate from the recent past, among other cases.

Which populations have the oldest mixing in the dataset?

The group with the longest time since admixture is detected are the Kalash from Pakistan, with an ancient inferred event prior to 206BCE, involving mixing between a more European and West Asian group, and a more Central/South Asian group (there may also be a contribution from people carrying DNA shared with modern-day East Asians, but we are less certain about this). Some Kalash believe they are descended from the army of Alexander the Great – our date does not rule this out but the date range also allows for many other possibilities. Other Pakistani groups also show equally ancient events with similar (but slightly less European-like) groups mixing. Other very ancient events involving

completely different sources are also seen, for example in Ethiopians, and Russian people. All these populations have additional mixing events more recently. The geographical isolation of the Kalash might explain why they don't show any recent event signal.

How do your results on the Mongol expansion relate to previous analyses?

We are not the first authors to assert that Genghis Khan and his armies had a genetic impact across Asia, although previous analyses were based on more limited genetic data. Over ten years ago, comparisons of Y chromosome lineages across Europe and Asia showed that a large number of Asian men shared a common Y chromosome (belong to "haplogroup" C*(C3c); Zerjal et al 2003). The breadth of the distribution of this haplogroup, and the author's extrapolation that roughly 0.5% of the world's population carried the same Y chromosome, together with age estimates (based on microsatellite diversity) placing the haplogroup at <34 generations old, led to the idea that Genghis Khan himself might have left a lasting genetic impact across Asia. More recently, a study using genome-wide data with different methods and genetic markers but on a similar (but smaller) set of populations to those used in our paper (Patterson et al. 2012), found evidence of admixture in the Uyghurs, dating to the time of Genghis Khan. As well as the Uyghurs, we found evidence of this Mongolian expansion in a further 6 populations, all with similar dates, and sometimes much further west. These populations approximately span the maximum spread of the Mongol empire. There are many central Asian and Eurasian populations in our analysis that don't show evidence of Mongolian admixture, implying that most Asian populations were *not* affected by this expansion. Taken together, we believe that there is now strong evidence that this event had a major impact on many Eurasian populations.

What explains the Eastern Europe results?

We give a more detailed discussion on the webpage, but briefly we find evidence that three different source groups intermixed in each of six Eastern European populations. Because this event is complex, we're uncertain of the precise nature of the groups involved, but one carries more North east-Asian like DNA contributing up to 4.5% of ancestry, another is more Northern European-like, and the third is more South European-like. All events happen around the same time: the period 300-1000CE. Another analysis (Ralph and Coop 2013) based on the length of shared ancestry chunks between individuals living in the same region shows signals of strong sharing of DNA, perhaps caused by rapid expansion of a population at very roughly the same time. Nomads from the central Eurasian Steppe are known to have moved into Eastern Europe during this period, and during the same period, perhaps carrying at least some North east-Asian like genetic material. Because the "northern" group are identified to be "most like" Polish, Lithuanian or Belorussian people in our analysis, we speculate the Northern European-like source may be Slavic speaking migrants, and the mixing date overlaps the period of the Slavic expansion across Eastern Europe.

Are there interesting signals in the data that need to be investigated further?

There are a number of populations that show admixture events that are not straightforward enough to be categorized by our current analysis. For example, the French show an event involving Northern and Southern European and North African populations dating to 1085 years ago plus or minus 300 years.

However, according to the automated quantitative criterion we developed for characterizing admixture events, this event is characterized as "uncertain". One reason for this is that several of the "coancestry curves" (see below), for example those involving the Moroccan population, show a steeper slope at short genetic distances than expected given an admixture event 1000 years ago. This suggests that there was perhaps at least one earlier admixture event that has left traces in the data. Other signals, such as those in a sample from the Orkney islands (Orcadians) are interesting but do not quite meet our measure of evidence of admixture. Here, as in many other parts of the world, larger numbers of samples from the French, Orcadians and from surrounding populations should allow this signal to be refined.

What follow up projects are you doing?

The absence of signals in some parts of Northern Europe and China likely reflects a lack of power to detect subtle signals – with e.g. the historically evidenced major historical migrations into the UK probably chiefly originating from nearby Northern European populations – based on our sample sizes, rather than an absence of events. One project that we are excited to be involved in is the Peopling of the British Isles project, studying a far greater number of individuals from the UK using related methods, and where results so far look very interesting. There are also many, many additional events worldwide waiting to be unearthed. Ongoing projects include more detailed exploration of Eurasian populations as well as others from the Caribbean, Ethiopia, and the Americas.

Can your methods be applied to other species than humans?

Yes, certainly, although in some species, the absence of a good recombination map may present a challenge.

QUESTIONS ABOUT THE METHOD: Statistical inference from genomic data using FineSTRUCTURE and GLOBETROTTER

What does fineSTRUCTURE do and how is it used here?

FineSTRUCTURE is statistical software (freely available at www.paintmychromosomes.com) that clusters sampled individuals into distinct groups based solely on their genetic similarities with other individuals. In particular, fineSTRUCTURE will cluster together individuals that are more genetically similar relative to all other sampled individuals. We applied fineSTRUCTURE here in order to identify individuals within each of our 95 sampled groups who appeared to be genetically different from the majority of individuals within that group. These genetically different individuals were then removed before we analyzed that group, to avoid confusing the analysis. Typically 0-3 individuals were removed from each group (usually 0), with >3 removed in only three cases.

What is chromosome painting?

Imagine the DNA of each world-wide group is represented by a different color. The CHROMOPAINTER algorithm (also available at www.paintmychromosomes.com) takes a string of DNA from the beginning of the chromosome of an individual and colors it (i.e. "paints" it) according to the world-wide group of the chromosome in the sample that it most closely matches. Such a match

implies recent shared ancestry with that world-wide group. Each of our chromosomes is a mosaic, stitched together from the chromosomes of our ancestors. The CHROMOPAINTER algorithm also detects the point at which the closest match in the samples changes and paints the next string based on the identity of the new closest match. Hence each chromosome is painted using multiple colors. Because human variation is usually widely shared across populations, this painting does not match the underlying ancestry completely, but does give a hint – one way of thinking about this is that the painting is very "noisy". Our approach uses modeling to reduce this noise.

What is mixture modeling?

A typical individual is broken into ~30,000 strings by CHROMOPAINTER, meaning that we have ~30,000 pieces of information on his or her ancestry. All of the groups in the sample are painted with multiple colors, reflecting extensive sharing of genetic variation between human populations. The palette (or painting profile) of each group is the proportion of the stretches DNA that are painted using each color. Painting palettes differ between groups as illustrated below (Figure 3). Mixture modelling uses the palettes of different groups in order to help to characterize their admixture history.

A)

B)

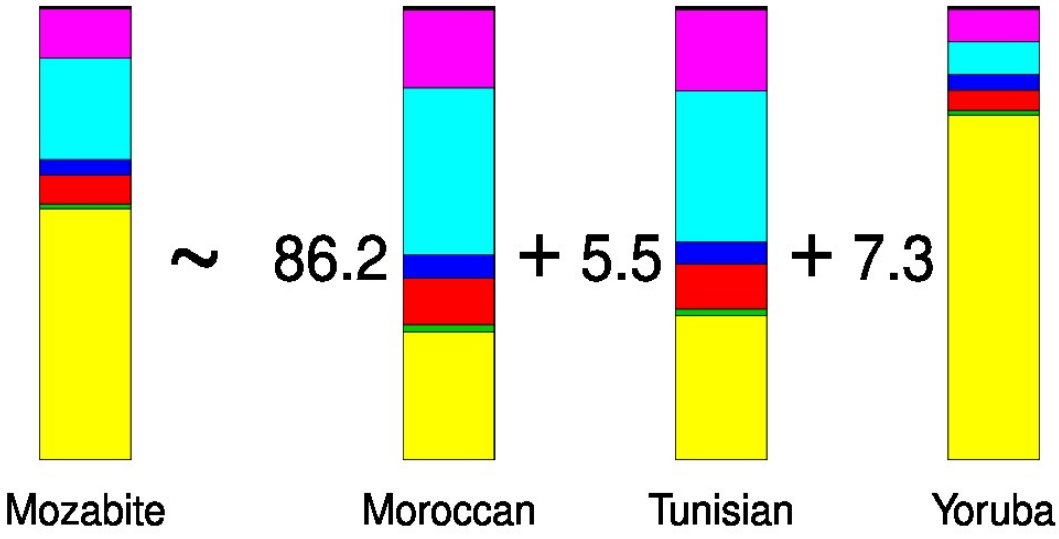

Figure 3. (A) painting of 20MB of DNA from a Mozabite chromosome, (B) the painting palette of the Mozabite group and the mixture modeling reconstruction of the palette as a mix of Moroccan, Tunisian and Yoruba palettes. Here the colours represent yellow, Africa; green, America; red, Central-South Asia; blue, East Asia; cyan, Europe; pink, Near East; black Oceania.

What are coancestry curves?
Coancestry curves capture information about the lengths of segments inherited from each original admixing source group. They do this by tabulating how rapidly the painting palette changes as genetic distance increases along the chromosome.

Because every ethnic group is painted with a multicoloured palette, it is never certain from which population any particular stretch of DNA comes from. However, the painting provides noisy information on the true ancestry of each stretch. For two different stretches of DNA (chunks) separated by some distance, they can then in turn give noisy information about whether this true ancestry changes between them, or stays the same. By summing information over large numbers of different stretches from each of the chromosomes of multiple individuals, it is often possible to extract clean curves that reflect the length of segments from the original admixing source groups (Figure 4). All such curves are available at www.admixturemap.paintmychromosomes.com.

How are the coancestry curves used to date admixture events?
Theory suggests that these coancestry curves should follow an exponential distribution with rate equal to number of generations ago that the admixture event occurred. We therefore estimate an exponential decay rate simultaneously to our full set of coancestry curves using standard statistical software. We also compare a model with two admixture times to a model with a single admixture time. If there are two or more admixture times, then the coancestry curves are predicted theoretically to have a mixture of exponential decay rates.

How do you convert your inferred dates of admixture in generations ago to a date in years?
Following current literature, we assume a generation corresponds to 28 years. We then convert generations (G) to years (Y) using the following formula: $Y = 1950 - 28 \times (G+1)$. This formula fixes 1950 as the arbitrary birth-year for our sampled individuals, which should not make a significant difference given the overall uncertainty in date inference.

Some of the coancestry curves slope upwards and some slope downwards, to different values on the Y-axis. What does this mean?
When exploring admixture in population X, a sampled group that has coancestry curves showing a clear exponential pattern suggests that sampled group is important for representing one of the original admixing source groups of population X. Theory predicts that if two different sampled groups have a coancestry curve that is *decreasing* with distance, these two groups should represent the same admixing source (e.g. Tunisian and Moroccans above). Conversely, if two different sampled groups have a coancestry curve that is *increasing* with distance, these two groups each represent a different admixing source (e.g. Tunisians and Yoruba above). This information, and the values at which curves meet the Y-axis, allow us with the mixture modeling to separate out the admixing sources, and are also used to improve the mixture modeling itself.

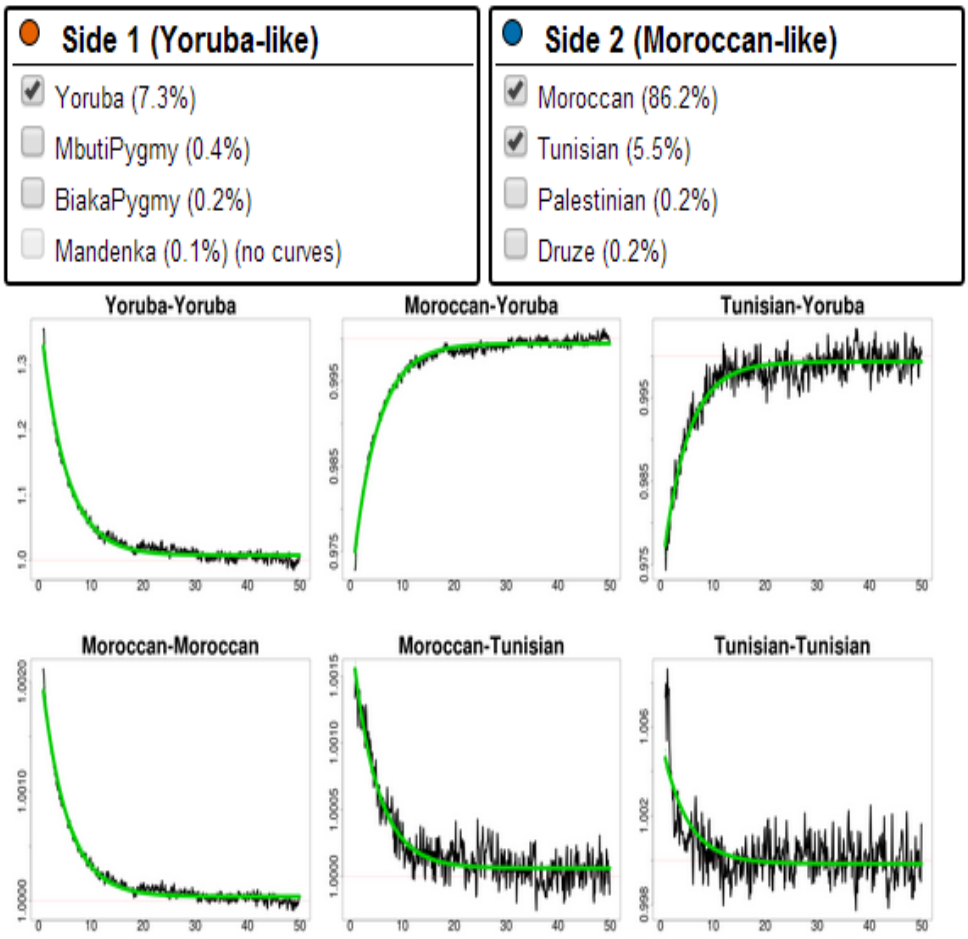

Figure 4. Coancestry curves for Mozabites. Shows coancestry as a function of genetic distance in centiMorgans. The curves were fit to a single admixture date of 1334CE. The possible departure for the "Tunisian" curves might suggest an older date – but is also consistent with random noise according to our testing. Taken from admixturemap.paintmychromsomes.com

How did you test your method?

We tested our method using 4700 whole-genome simulations from 80 different admixture events. Some of these were based on generating data from scratch according to realistic parameters for human evolution. Others were based on taking real chromosomes and generating artificial admixture events *in silico*. Some problems were easy (like finding recent African admixture in Europeans), others were hard (4,000 year old admixture between central Asian and European groups) and some were impossible (no admixture at all). Our method succeeded in not finding admixture events were there were none and generally did a very good job of reconstructing the admixture events we had simulated unless the problem was very tricky. Problems become trickier if (1) more than one admixture event took place in the ancestry of a sample, (2) the contribution of a source population was very small (below 5%), (3) if the admixture was old and (4) if the admixing sources were very similar. More details are presented in the supplement of our paper.

Is your analysis objective?

Our method is objective in that it uses genetic data – and only genetic data – to determine if, when, and where genetic admixture has occurred, without prior "guesses". In particular, we take a large and

diverse collection of samples (here from 95 world-wide groups) and automatically infer which samples have evidence of admixture, dating any identified admixture events and describing the DNA of the original admixing sources as mixtures of the DNA from our present-day sampled populations. Thresholds in the inference were obtained theoretically and/or using a large collection of simulations, which we also used to test the method.

We also did a number of “follow up” analyses, including some region specific analyses and although they use the identical approach, these analyses can be considered to be a bit less objective than the global analysis in that we used them to test specific hypotheses. Based on the results, we also identified 10 groups of events that we think likely reflect particular historic events. This process of interpretation in the light of anthropological, archaeological, historical and linguistic sources is necessarily more subjective. Some regions of the world, e.g. parts of the Americas, also have few or no samples, preventing our analyzing them.